Bioinformatics and its applications

Alla L Lapidus, Ph.D.

SPbAU, SPbSU,

St. Petersburg

Term Bioinformatics

Term **Bioinformatics** was invented by Paulien Hogeweg (Полина Хогевег) and Ben Hesper in 1970 as "the study of informatic processes in biotic systems".

Paulien Hogeweg is a Dutch theoretical biologist and complex systems researcher studying biological systems as dynamic information processing systems at many interconnected levels.

Definitions of what is Bioinformatics:

Bioinformatics is the use of IT in biotechnology for the data storage, data warehousing and analyzing the DNA sequences. In Bioinfomatics knowledge of many branches are required like biology, mathematics, computer science, laws of physics & chemistry, and of course sound knowledge of IT to analyze biotech data. Bioinforn
used to solve n

Bioinformatics is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful

The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related informa

Bioinformatics develop
of computer to analy
biology for example
ingredients and metab

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

http://www.bisti.nih.gov/CompuBioDef.pdf

My additions:

1. Bioinformatics is a SCIENCE

2. Not only to develop algorithms, store, retrieve, organize and analyze biological data but to CURATE data

Bioinformatics is being used in following fields:

- Microbial genome applications
- Molecular medicine
- Personalised medicine
- Preventative medicine
- Gene therapy
- Drug development
- Antibiotic resistance
- Evolutionary studies
- Waste cleanup
- Biotechnology
- Climate change Studies
- Alternative energy sources
- Crop improvement
- Forensic analysis
- Bio-weapon creation
- Insect resistance
- Improve nutritional quality
- Development of Drought resistant varieties
- Vetinary Science

Sequencing projects

DNA QC

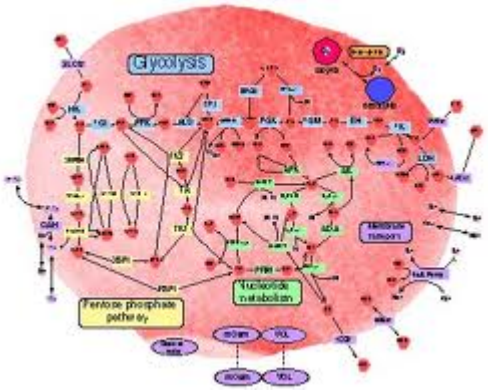
lib QC

Run QC

Data QC

DNA

Libraries



Data analysis

Results Interpretation

Data applications

LIMS - Lab Information Management Software

Microbial genome applications

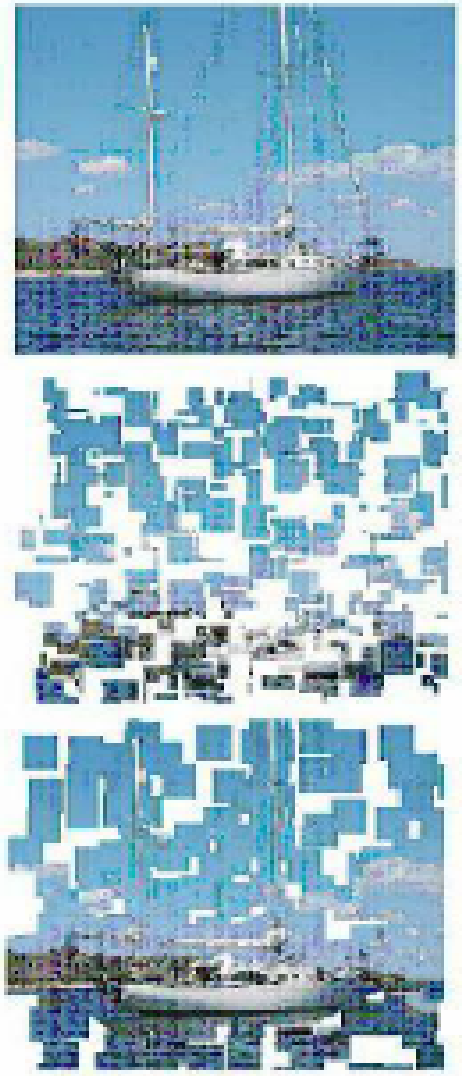
- Genome assembly
- Re-sequencing
- Comparative analysis
- Evolutionary studies
- Antibiotic resistance
- Waste cleanup
- Biotechnology

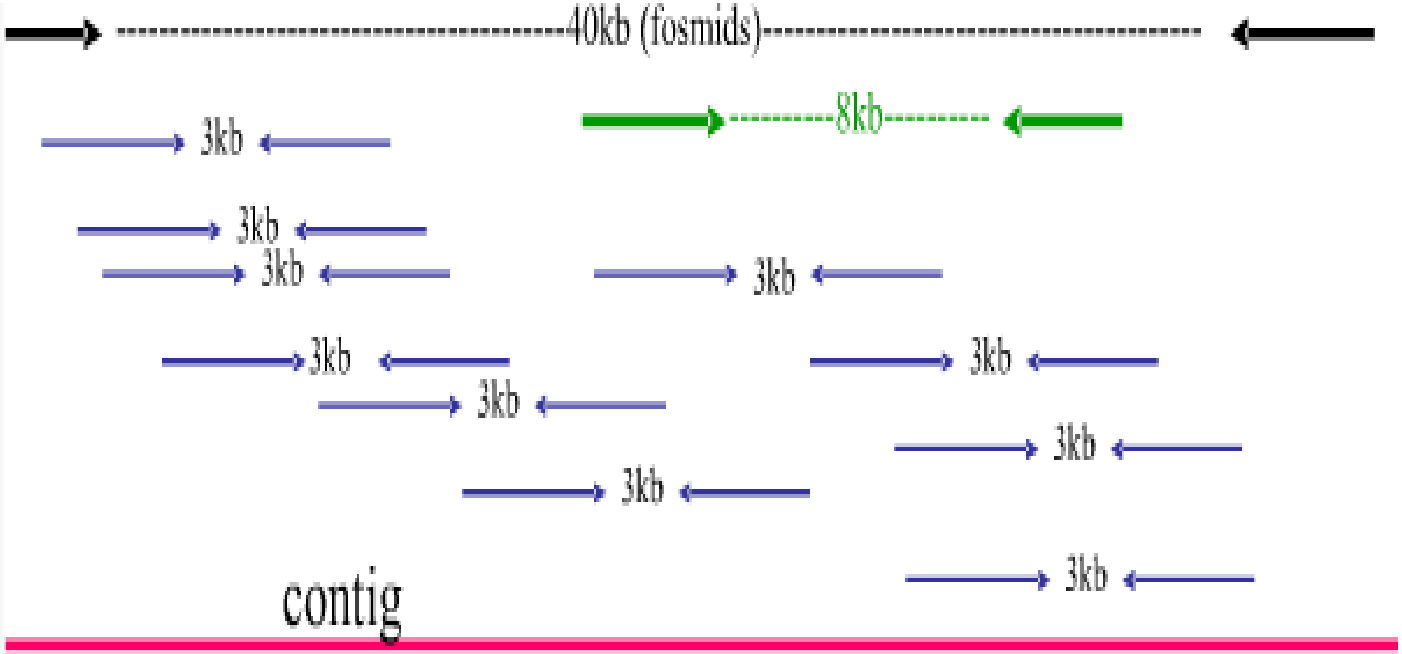
Genome Assembly

- Genome assembly is a very complex computational problem due to enormous amount of data to put together and some other reasons reasons.

- Ideally an assembly program should produce one contig for every chromosome of the genome being sequenced. But because of the complex nature of the genomes, the ideal conditions just never possible, thus leading to gaps in the genome.

De Novo assembly - puzzle without the picture



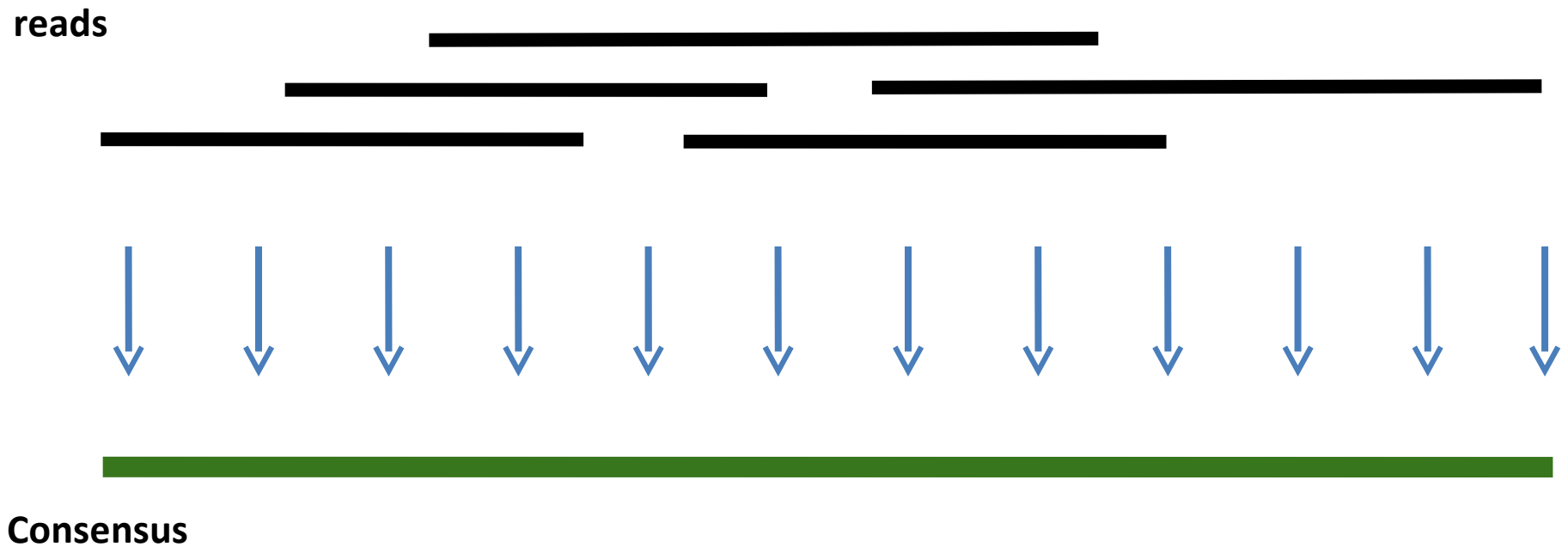


Assembly Challenges

- Presence of **repeats**. Repeats are identical sequences that occur in the genome in different locations and are often seen in varying lengths and in the multiple copies. There are several types of repeats: tandem repeats or interspersed repeats. The read's originating from different copies of the repeat appear identical to the assembler, causing errors in the assembly.
- **Contaminants** in samples (eg. from Bacteria or Human).
- **PCR artefacts (*eg.* Chimeras** and **Mutations)**
- **Sequencing errors**, such as **"*Homopolymer*" errors** – when eg. 2+ run of same base.
- **MID**'s (multiplex indexes), **primers/adapters** still in the raw reads.
- **polyploid** genomes

Assembly algorithms

Overlap-Layout-Consensus - Find overlaps between all reads

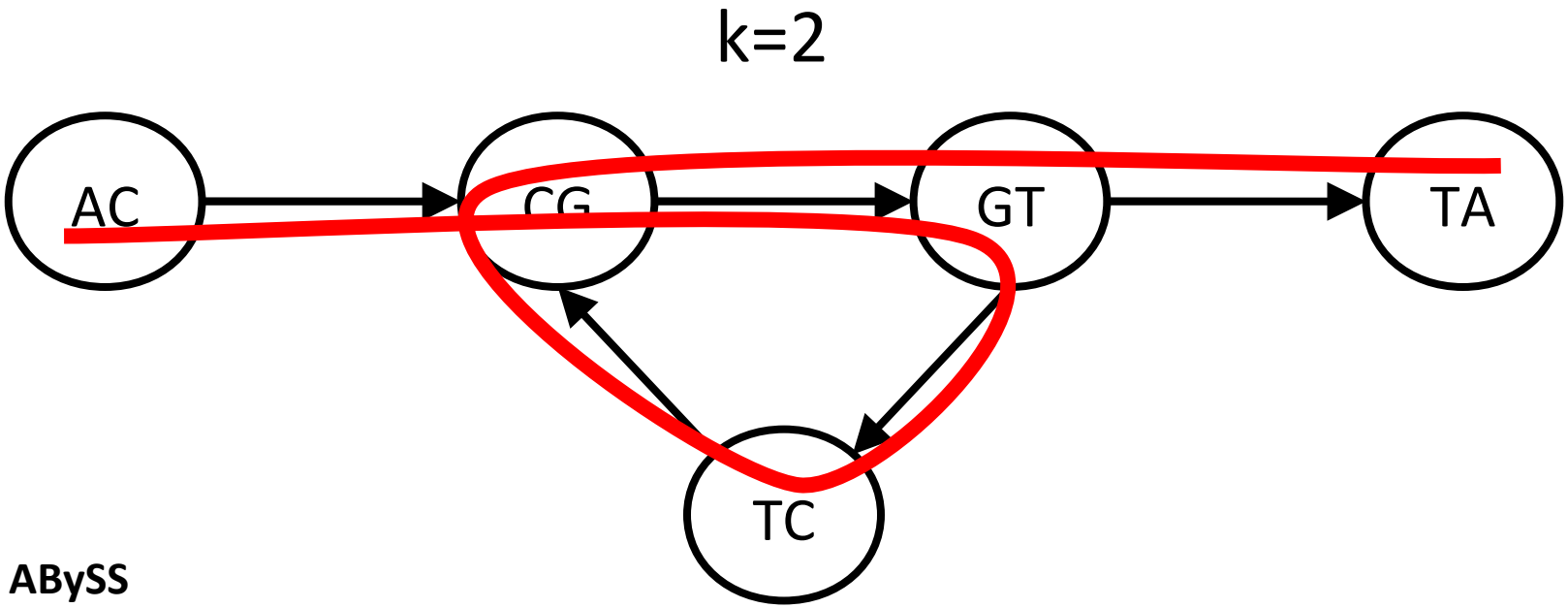


Problems caused by new sequencing technologies:

- Hard to find overlaps between short reads
- Impossible to scale up

De Bruijn graph

ACGTCGTA



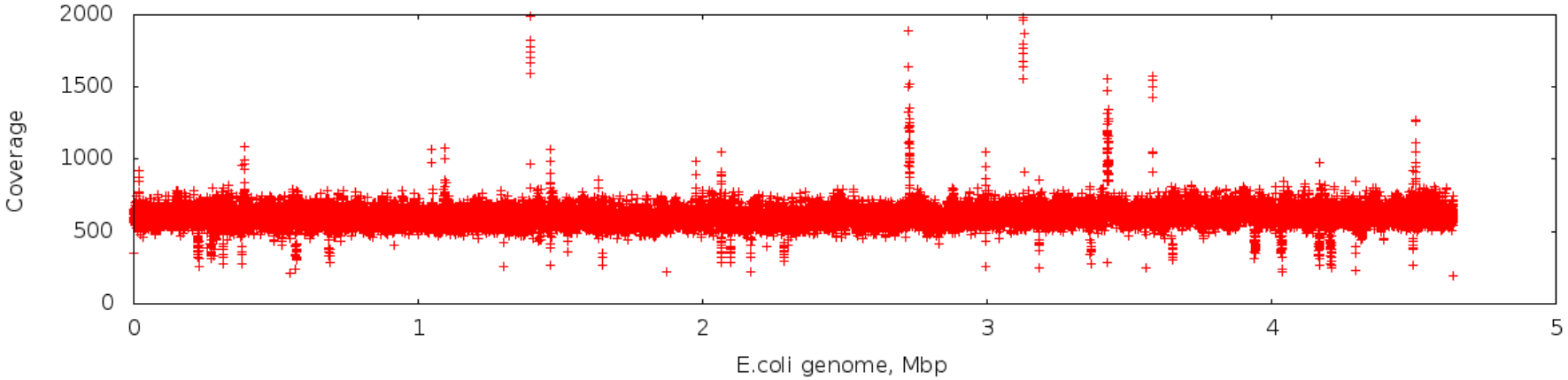
- **ABySS**
- **ALLPATHS-LG**
- **EULER**
- **IDBA**
- **Velvet**

Modified from Andrey Prjibelski

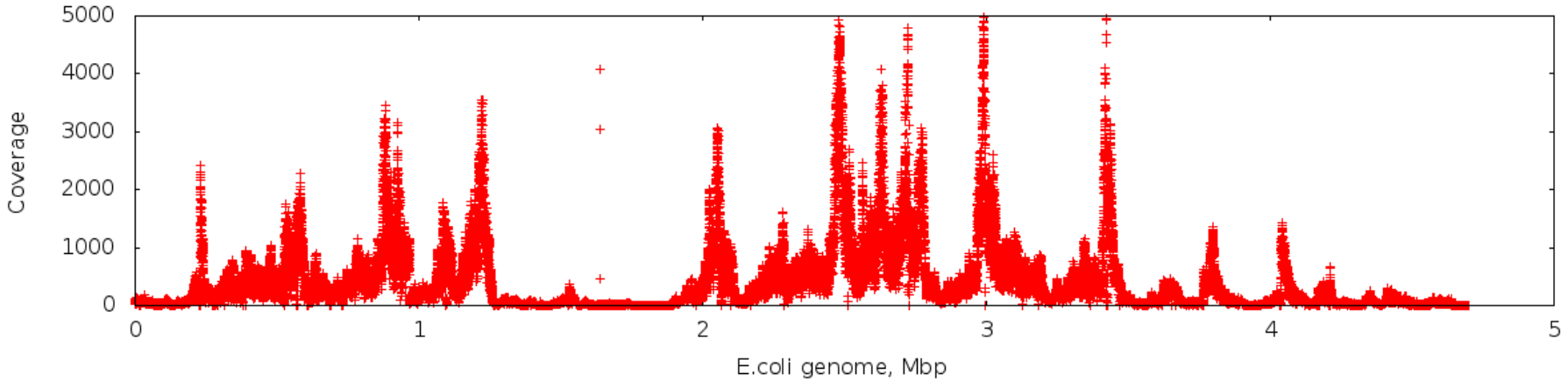
Single-cell dataset

- *E. coli* isolate dataset

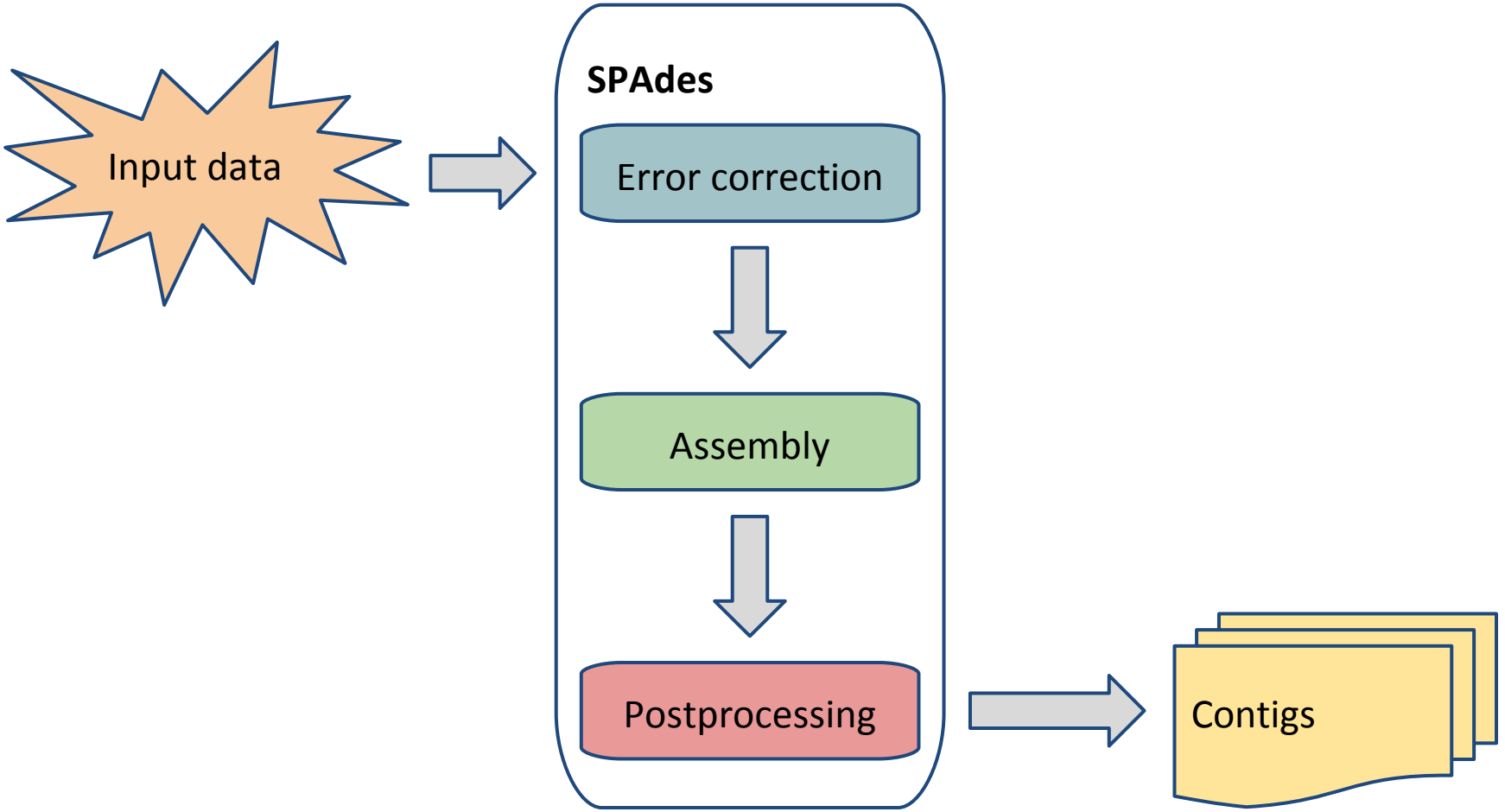
- IDBA-UD
- SPAdes
- Velvet

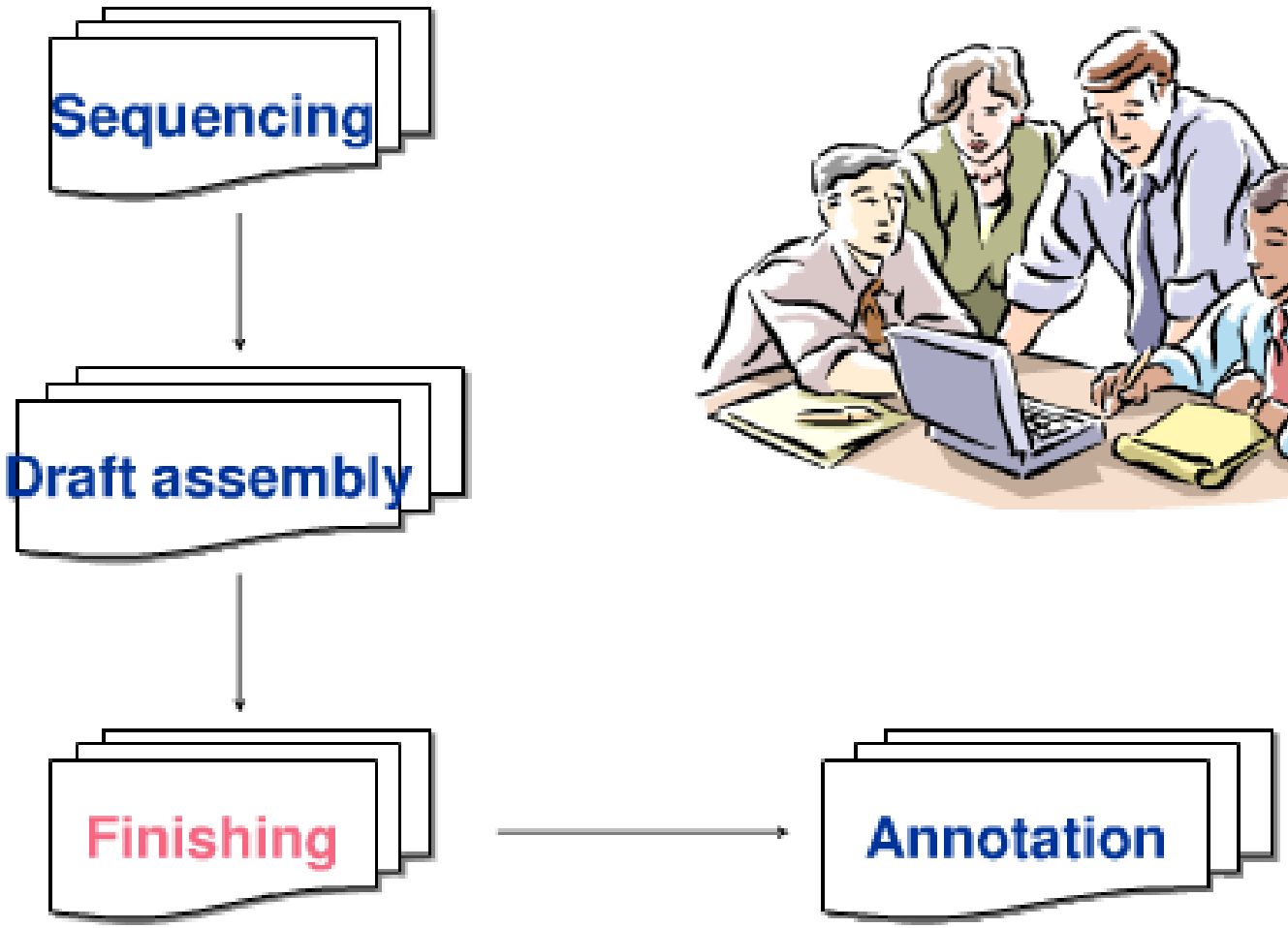


- *E.coli* single-cell dataset



SPAdes pipeline




Sequencing
Draft assembly
Finishing
Annotation

Gene Prediction and Genome Annotation

Based on similarity to known genes – blastX (NCBI)

Gene finding programs

- **Glimmer** – for most procaryotic genomes
- **GenMark** – for both procaryotic genomes and eucaryotic genomes

INTEGRATED MICROBIAL GENOMES

IMG Content

	Datasets
Bacteria	6120
Archaea	248
Eukarya	183
Plasmids	1193
Viruses	2809
Genome Fragments	579
Total Datasets	11132
GEBA	245

Last updated: 2013-07-05

IMG 4.0 is dedicated to the memory of our colleague, Iain Anderson

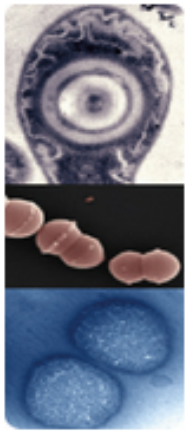
Genome by Metadata
Project Map
Content History
System Requirements
About IMG
FAQ

Hands on training available at the

Microbial Genomics & Metagenomics Workshop

The **Integrated Microbial Genomes (IMG)** system (Nucleic Acids Research, Vol 40, 2012) serves as a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life in a uniquely integrated context. Plasmids that are not part of a specific microbial genome sequencing project and phage genomes are also included into IMG in order to increase its genomic context for comparative analysis.

For details, see IMG Release Notes (Dec. 12, 2012), in particular the workspace and background computation capabilities available to IMG registered users.



Count	Total
DNA, number of bases	60,476,662,654
Total Genes	25,395,838
Total Genomes	11,132

IMG Statistics **IMG ER Account Request**

All Genomes

Genome Count

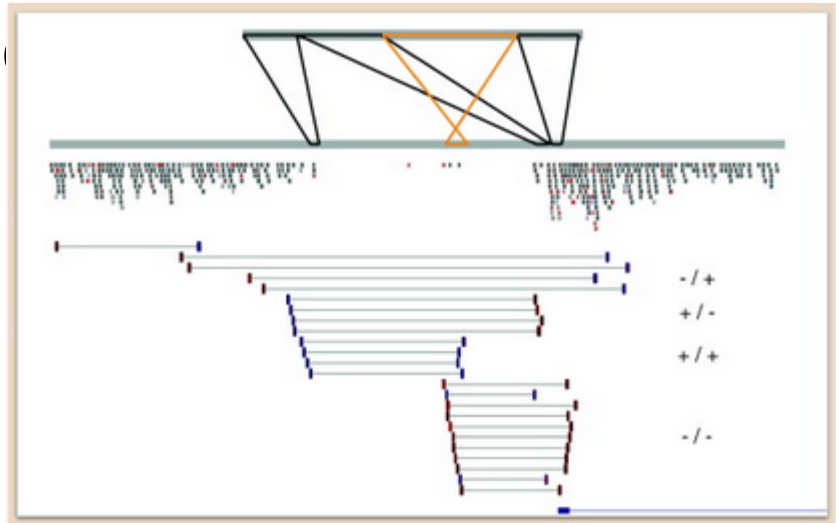
Status	Bacteria	Archaea	Eukaryota	Plasmids	Viruses	Genome Fragments	Total
Finished	2131	154	37	1190	2809	579	6900
Draft	2407	28	146	3	0	0	2584
Permanent Draft	1582	66	0	0	0	0	1648
Total	6120	248	183	1193	2809	579	11132

Re-sequencing

Projects aimed at characterizing the genetic variations of species or populations

Resequencing of bacterial and archaeal isolates etc is possible if reference genomes are available

This approach can help to better understand bacterial community structure, gene function in bacteria under selective pr



Climate change Studies

Increasing levels of carbon dioxide emission are thought to contribute to global climate change.

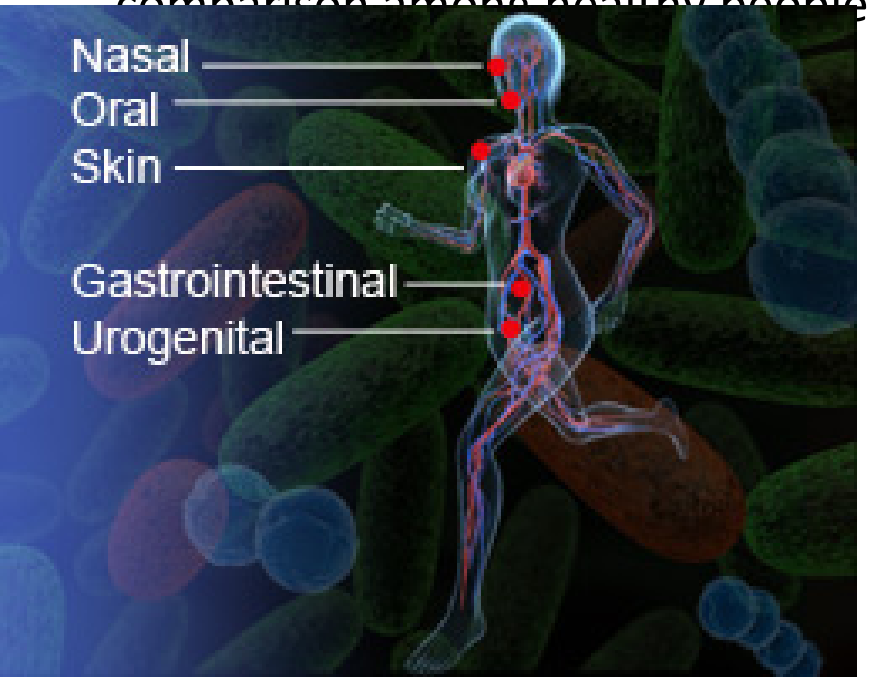
One way to decrease atmospheric carbon dioxide is to study the genomes of microbes that use carbone dioxidet as their sole carbon source

Human microbiome

MetaHIT - Europe
Human Microbiome Project –US

The human microbiome includes viruses, fungi and bacteria, their genes and their environmental interactions, and is known to influence human physiology.

There's very broad variation in these bacteria in different people and that severely limits our ability to create a "normal" microflora profile for comparison among healthy people and those with any kind of health issues.



Children with **autism** harbor significantly fewer types of gut bacteria than those who are not affected by the disorder, researchers have found.

Prevotella species were most dramatically reduced among samples from autistic children—especially *P. copri*. (helps the breakdown of protein and carbohydrate foods)

Bioinformatics combining biology with computer science

- it can explore the causes of diseases at the molecular level
- explain the phenomena of the diseases on the gene/pathway level
- make use of computer techniques (data mining, machine learning etc), to analyze and interpret data faster
- to enhance the accuracy of the results

Reduce the cost and time of drug discovery

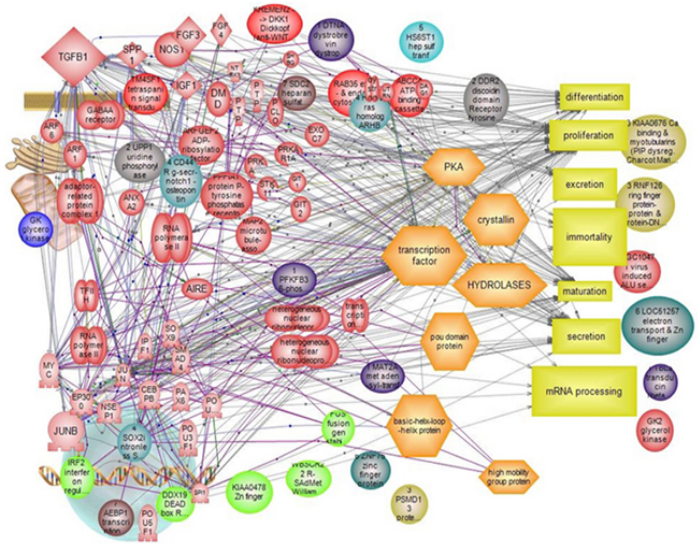
To improve drug discovery we need to discover (read "develop") efficient bioinformatics algorithms and approaches for

target identification

target validation

lead identification

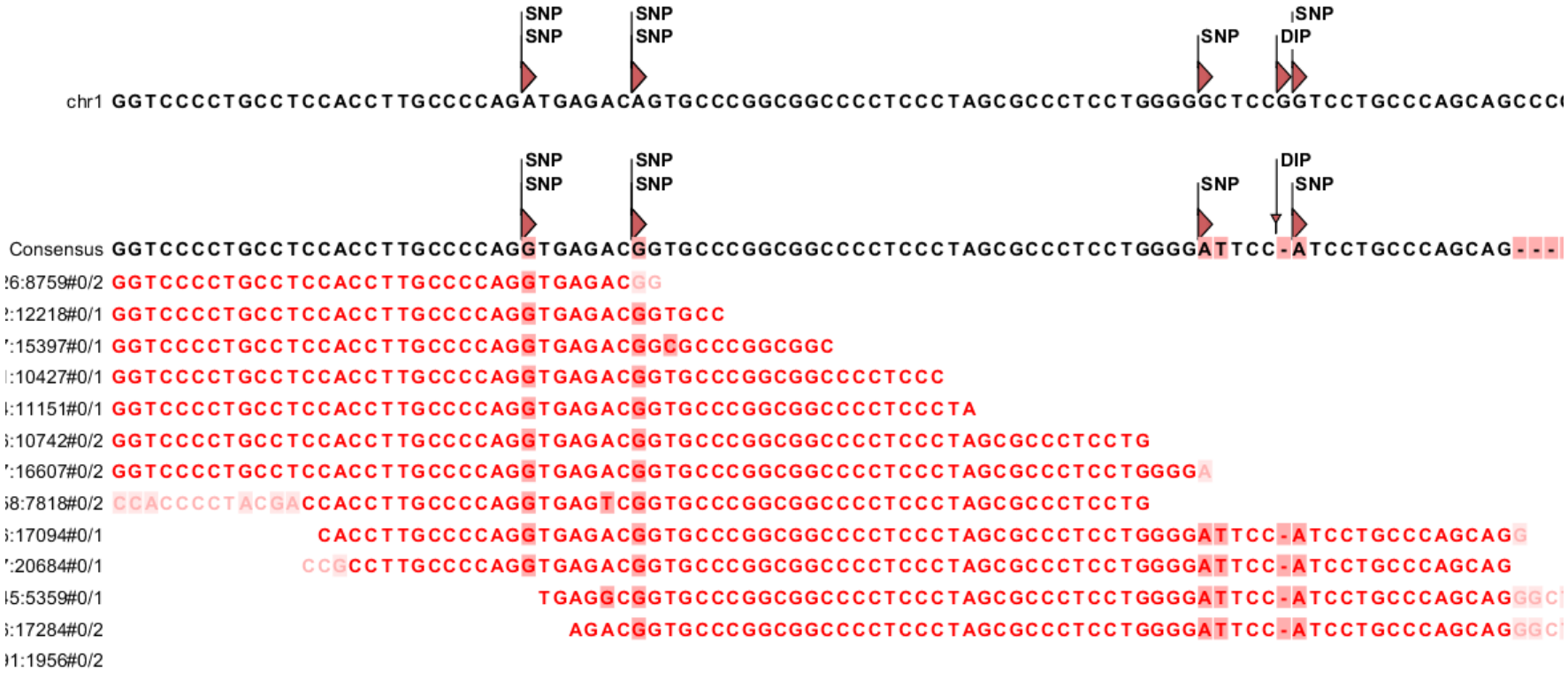
lead optimization



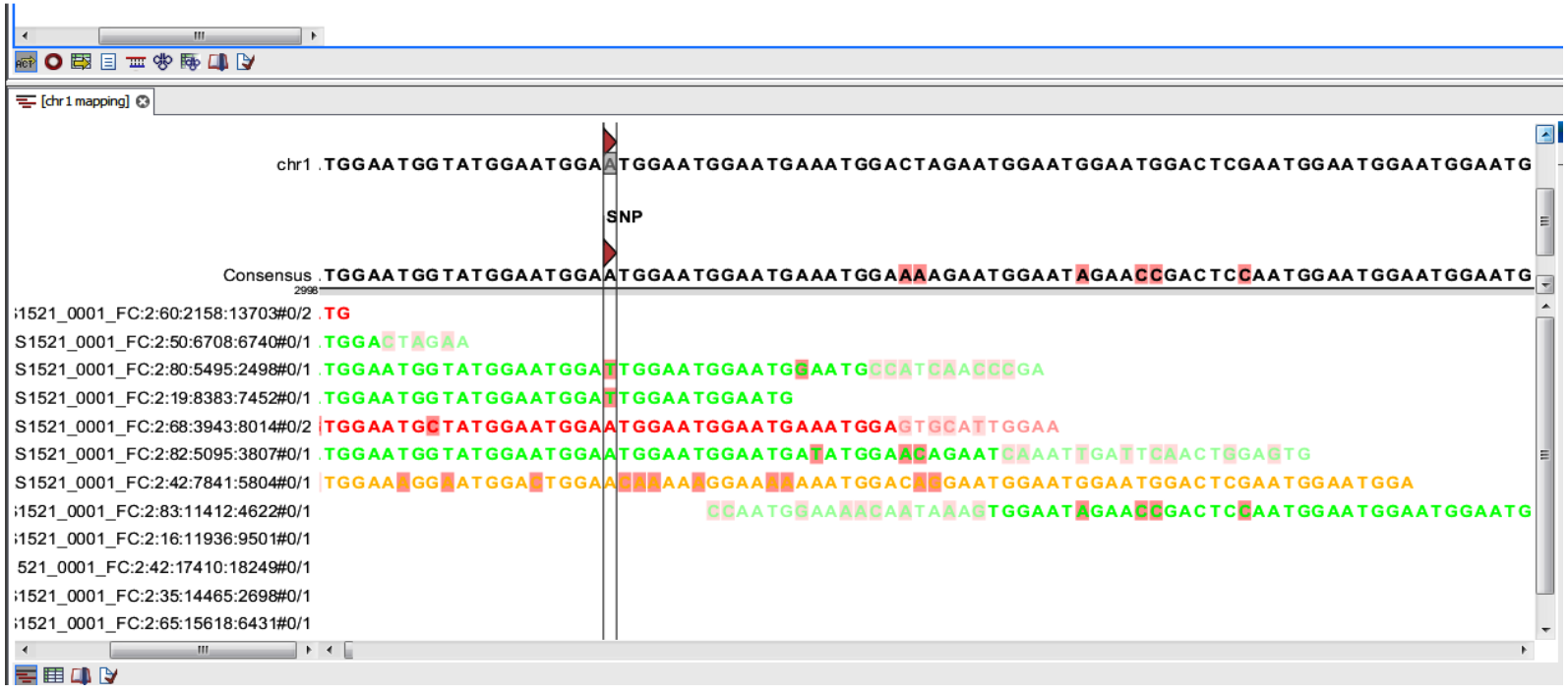
Advantages of detecting mutations with next-generation sequencing

- High throughput
  - Test many genes at once
- Systematic, unbiased mutation detection
  - All mutation types
    - Single nucleotide variants (SNV), copy number variation (CNV)-insertions, deletions and translocations
- Digital readout of mutation frequency
  - Easier to detect and quantify mutations in a heterogeneous sample
- Cost effective **precision** medicine
  - "Right drug at right dose to the right patient at the right time"

Homozygous SNPs and indel



Poor alignment



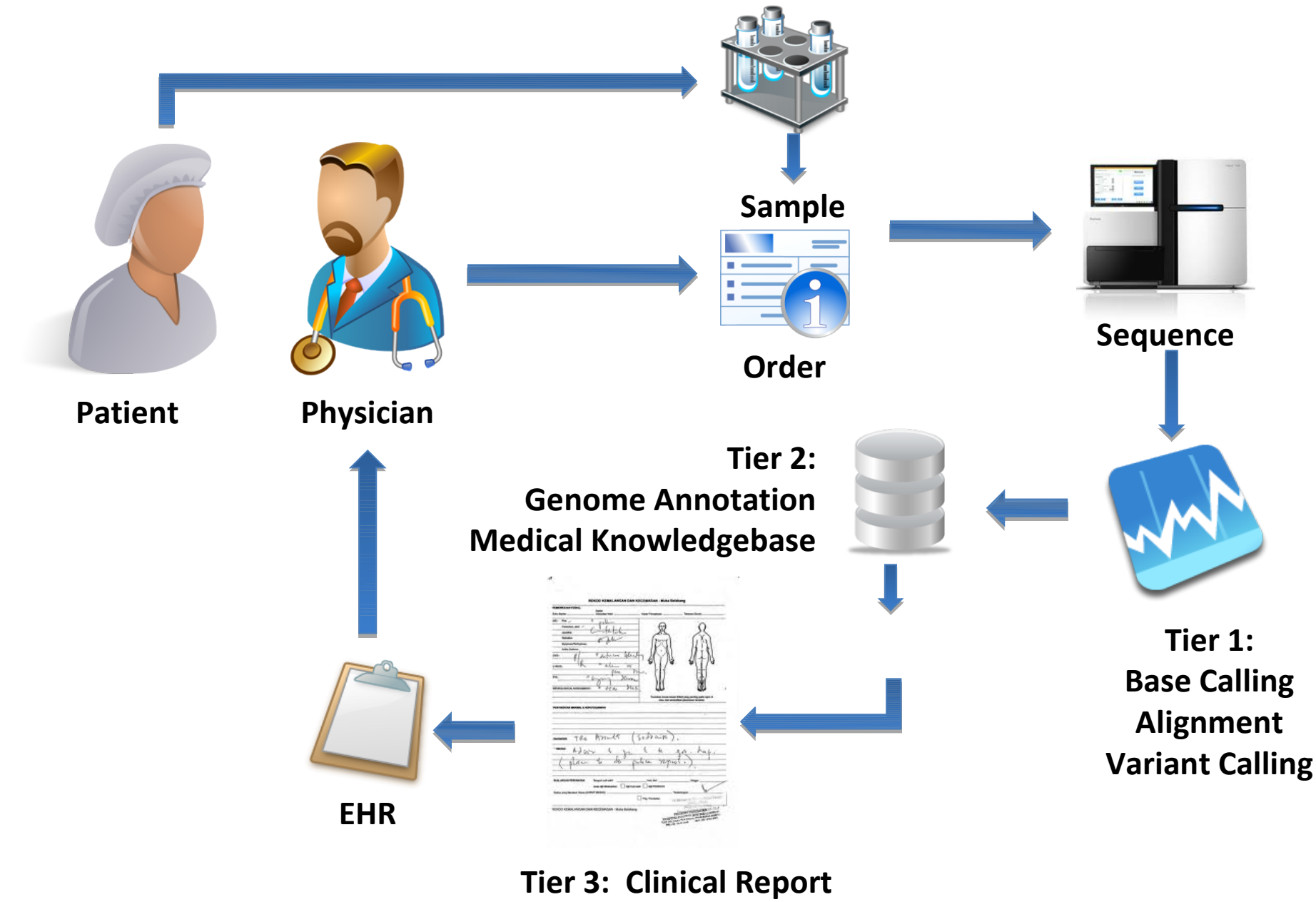
Missed SNP?


chrGCAGAGGCCAAGCCAGAGGTTCCAGGCTTAAACCCCAGCCCTGCCCTGCCCAGTCCA
ConsensuGCAGAGGCCAAGCCAGAGGTTCCAGGCTTAAACCCCAGCCCTGCCCTGCCCAGTCCA
3866:4795#0/GCA
1407:2153#0/
4308:3912#0/GCAGAGGCCAAGCCAGAGGTTCCAGGCTTAAA
9914:7870#0/ GCCAAGCCAGAGGTTCCAGGCTTAAACCCCAGCCCTGCCCTGCCCAGTCCA
5555:2114#0/ CAAGCCAGAGGTTCCAGGCTTAAACCCCAGCCCTGCCCTGCCCAGTCCA
579:16341#0/ CCAGGCTTAACCCCCAGCCCTGCCCTGCCCAGTCCA
6944:9734#0/ GGCTTAACCCCCAGCCCTGCCCTGCCCAGTCCA
408:13945#0/

Bioinformatics and Health Informatics

If bioinformatics is the study of the flow of information in biological sciences, Health Informatics is the study of the information in patient care

Medicine: Informatics pipeline workflow



Huge need in bioinformatics tools

Simple pipelines/protocols and easy to read reports

Sample sequencing

Data Analysis

Patients treatment

Team work to set up cancer sequencing facility

statistician

statistician

molecular biologist

mathematician

data processing specialist

molecular oncologists (pathway analysis)

Data analysis automation

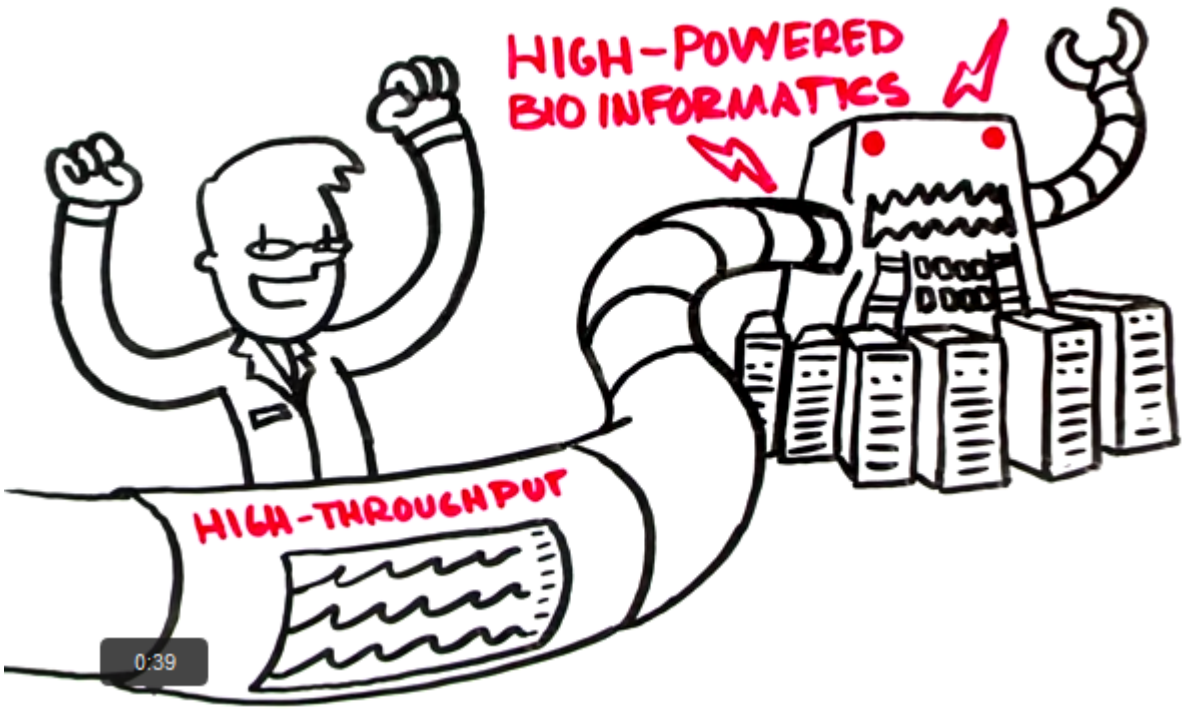
pathologists

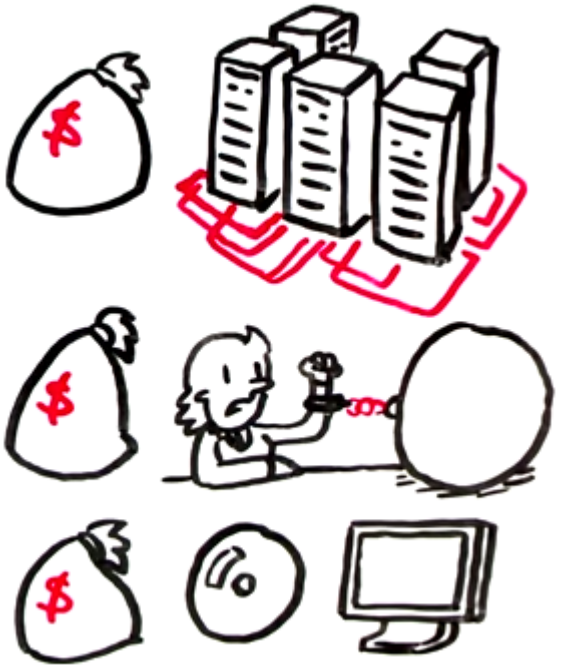
data bases specialist

Oncologists treatment selection

IT infrastructure (storage, updates etc)

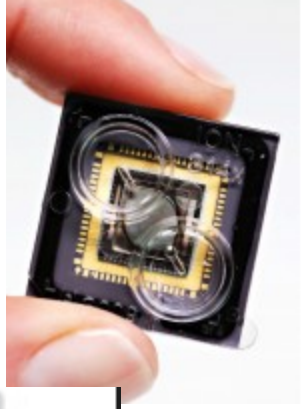
User friendly interfaces





http://www.youtube.com

Ion Torrent: Torrent Suite Software

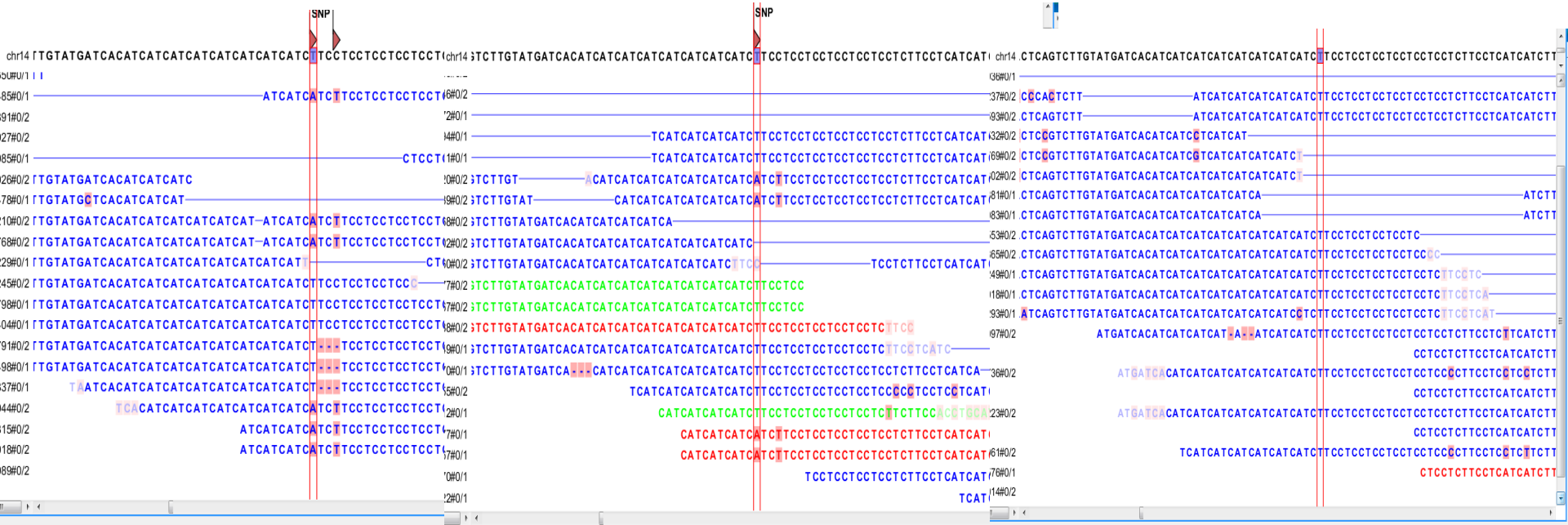


HOMEZ_22814666

S2

S4

S3

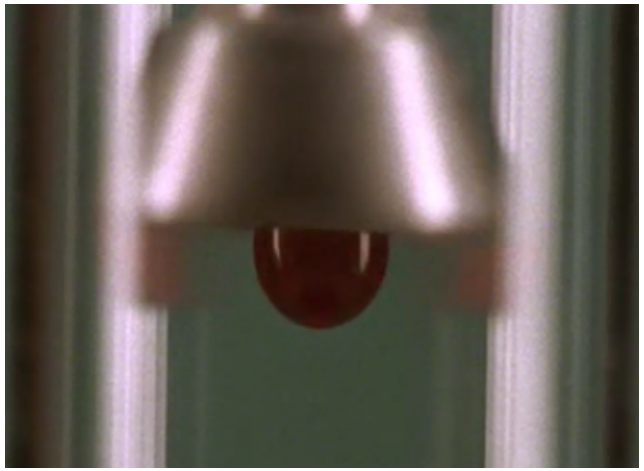


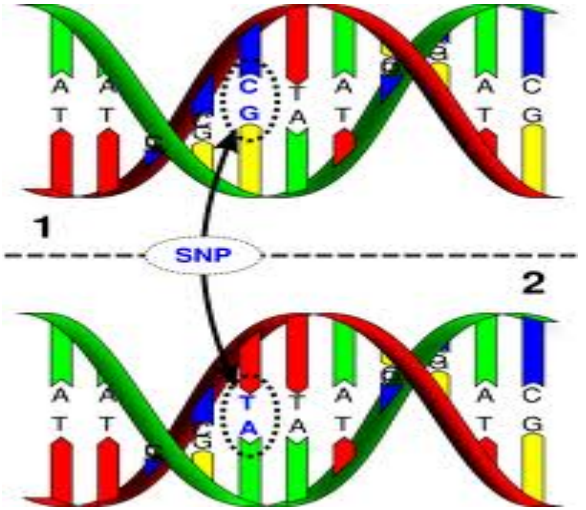
S8

S5

S7

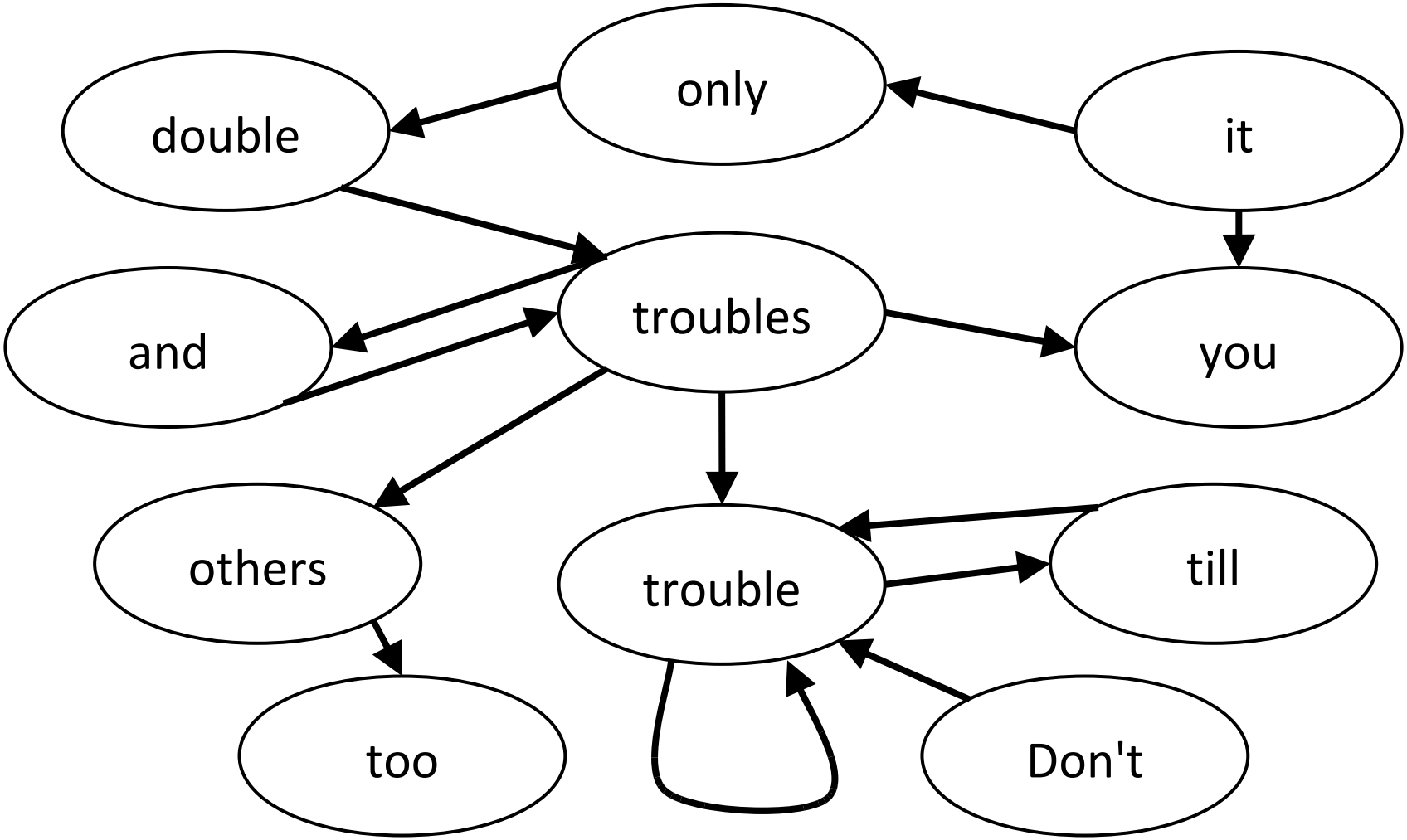
Each baby to be sequenced at birth: personal reference





“GATTACA”, 1997

Funny De Bruijn graph



Modified from Andrey Prjibelski

THANK YOU!